5-1-2010

# An Evaluation of the Convergent Validity of Multi-Source Feedback with Situational Assessment of Leadership - Student Assessment (SALSA©)

Heather Stroupe
*Western Kentucky University*, hstroupe@live.com

# AN EVALUATION OF THE CONVERGENT VALIDITY OF MULTI-SOURCE FEEDBACK WITH SITUATIONAL ASSESSMENT OF LEADERSHIP – STUDENT ASSESSMENT (SALSA©)

A Thesis
Presented to
The Faculty of the Department of Industrial/Organizational Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts

By
Heather Stroupe

May 2010

AN EVALUATION OF THE CONVERGENT VALIDITY OF

MULTI-SOURCE FEEDBACK WITH SITUATIONAL ASSESSMENT OF

LEADERSHIP – STUDENT ASSESSMENT (SALSA©)

Date Recommended: __April 23, 2010____

__Elizabeth Shoenfelt_____
Director of Thesis

__Jacqueline Pope-Tarrence_____

__Reagan Brown_____

_____
Dean, Graduate Studies and Research Date

TABLE OF CONTENTS

AN EVALUATION OF THE CONVERGENT VALIDITY OF MULTI-SOURCE FEEDBACK WITH SITUATIONAL ASSESSMENT OF LEADERSHIP – STUDENT ASSESSMENT (SALSA©)

Heather Stroupe May 2010 41 Pages

Directed by: Drs. Elizabeth L. Shoenfelt, Reagan Brown, Jacqueline Pope-Tarrence

Department of Psychology Western Kentucky University

The current study assessed the convergent validity of the Situational Assessment of Leadership – Student Assessment (SALSA©), a situational judgment test (SJT), with multi-source ratings. The SALSA© was administered to ROTC cadets via Blackboard; multi-source ratings, which paralleled the leadership dimensions of the SALSA©, were administered via paper. Each cadet completed the SALSA© and was rated by 10 peers, his/herself, and at least one cadre (superior). SALSA© scores were not correlated with any of the corresponding dimensions on multi-source ratings, with one exception. Cadre ratings of Consideration/Team Skills were positively correlated with SALSA© scores on the same dimension. This finding suggests that the multi-source ratings and the SALSA© are not measuring the same leadership construct. Self-ratings were significantly higher than peer or cadre ratings. Senior ROTC cadets scored significantly higher on SALSA© than did Junior ROTC cadets. Future research should focus on differences between autocratic styles of leadership and democratic styles of leadership and whether different SJTs are needed to measure each style.

## An Evaluation of the Convergent Validity of Multi-Source Feedback with Situational Assessment of Leadership – Student Assessment (SALSA©)

Shoenfelt (2008) developed the Situational Assessment of Leadership – Student Assessment (SALSA©), a situational judgment test (SJT) of leadership, based on a content-validation approach. The current study assessed the convergent validity of the recently developed SJT and multi-source feedback. That is, Army ROTC cadets completed SALSA©. Their test scores were correlated with self-, peer-, and Army cadre ratings.

In the following section, I review the literature on situational judgment tests (i.e., what they are, why they are used, how they are developed, etc.), multi-source feedback, the SJT in the current study (i.e., inferences from evaluations thus far), and how multi-source feedback data will be used to further validate the current SJT. First, I present an overview of SJTs.

### *Overview of Situational Judgment Tests*

Situational judgment tests are instruments used to predict future performance of employees. SJT test items consist of hypothetical situations that are likely to occur on the job. Response options are actions that range from good to poor performance. Test takers choose the response they would or should take in the given the situation; their response is intended to be indicative of their future behavior. This form of test often is used by employers to make hiring decisions, but is not a new technique. In World War II, situational civil service examinations were given to military applicants before admittance.

Since then, SJTs have grown into a common method used to aid in the hiring process (Lievens, Peeters, & Schollaert, 2008).

The development of SJTs usually follows Motowidlo, Dunnette, and Carter's (1990) approach. Typically, subject matter experts (SMEs) are asked to write brief scenarios based on critical incidents that will then be used on the SJT as a scenario or test item. The next step is to create response options for each of the scenarios. Typically, a new group of SMEs will read the scenarios and write descriptions of how they would respond. Once SMEs have generated alternative response options, a group of experienced SMEs with knowledge of the targeted domain (e.g., a tenured group of business executives) rate the response options to calibrate them and determine the correct or best response. Finally, a scoring key is determined a priori based upon these responses. The SJT development procedure explained by Motowildo et al. is not rigid; there are many variations that provide similar SJT products. For example, a developer could choose to have SMEs create potential or likely scenarios based on their knowledge of the domain versus creating scenarios based on past experience. Another variation may be to have the same group of SMEs create scenarios and multiple responses that range from a positive response, to a neutral response, to a negative response, and have a second SME group judge the effectiveness of the responses.

SJTs have grown in popularity because of their sound psychometric properties. In general, they broaden the criterion domain, have less adverse impact than strictly cognitive measures, and more face validity than cognitive measures (Lievens, Buyse, & Sackett, 2005). SJTs are useful for predicting job performance for several reasons (Motowidlo et al., 1990). The first explanation is the behavioral consistency principle,

which states that past behavior predicts future behavior (Lievens et al., 2008). As SJTs measure what a test-taker reportedly would do, his/her future behavior will likely correspond to his/her SJT response. Additionally, Lievens et al. (2008) noted that SJTs measure "applicants' intentions and goals," as well as do other useful predictors of job performance such as cognitive ability tasks or personality measures (p. 432).

Although SJTs can be designed to assess a variety of constructs, some constructs have received more attention than others. For example, Weekley and Jones (1999) conducted two studies of mid-level retail associates, and asked them to complete several measurements consisting of biodata information, cognitive ability ratings, performance ratings, and experience ratings. The data indicated that SJTs were significantly related to cognitive ability (weighted average $r = .45$), performance (weighted average $r = .19$), and experience (weighted average $r = .20$; Weekley & Jones). The results of the two studies were not consistent, however. The second study demonstrated that SJT scores did not fully mediate the relationship between cognitive ability and experience; that is, cognitive ability became less predictive as experience increased. The authors suggested that this research confirms that STJs should be viewed as a method, not a construct.

Weekley and Ployhart (2005) found similar results to those of Weekley and Jones (1999), but concluded there were additional correlates of SJTs. Data collected included measures of cognitive ability, training experience, an SJT, five-factor model inventory, and supervisor performance ratings. Weekley and Ployhart found significant correlations between the SJT and other predictors including job tenure ($r = .13$), cognitive ability ($r =$

.36), GPA ($r$ = .21), performance ($r$ = .22), conscientiousness ($r$ = .13), emotional stability ($r$ = .17), extroversion ($r$ = .14), and training experience ($r$ = .12).

McDaniel, Hartman, Whetzel, and Grubb's (2007) and Lievens et al.'s (2008) meta-analyses indicated that SJTs provide incremental validity (.03 to .08) over both cognitive measures and personality measures. Weekley and Ployhart (2005) found that their SJT showed incremental validity over cognitive ability (i.e., general mental ability and GPA), personality (i.e., The Big Five), and experience measures (i.e., general work experience). McDaniel et al. cautioned that while SJTs can have incremental validity over some predictors, there are scenarios where there will be near zero incremental validity. For example, McDaniel et al. discovered that adding a SJT to an existing battery, consisting of a cognitive ability measure and a Big Five test, decreased incremental validity (.01 and .02). Similarly, when the existing battery consisted of a cognitive ability measure and a SJT, observed incremental validity was a mere .01 and .03 with the addition of a Big Five test.

Despite the positive characteristics of SJTs, there are some concerns. Lievens et al. (2008) provided an empirical review of recent research. The review noted that internal consistency is often affected by the multidimensionality, length, and response instructions of SJTs, and that test-retest reliability is adequate. Results from their study indicated that internal consistency coefficients varied from .43 to .94 (Lievens et al.). Data indicated that longer SJTs tended to show higher internal consistency, and SJTs with directions asking participants 'to rate the effectiveness of each response' had the highest internal consistency (.73). Finally, because of the multidimensionality of SJTs, it may be best to

use test-retest reliability as a reliability measure for SJTs as factor analysis leads to low internal consistency (Lievens et al.).

A question of the fakability and coaching of SJT performance has been raised by other researchers. For example, faking may improve SJT scores from .08 to .89 *SD* (Lievens et al., 2008). However, faking has less of an impact on SJTs than on personality measures, and careful consideration of item transparency, cognitive loading, and response instructions moderates faking. On a positive note, because SJTs are situationally based, they tend to be exempt from coaching or practice effects.

Another concern with SJTs is whether Web-based SJTs are equivalent to paper-and-pencil tests in actual selection contexts. Ployhart and Ehrhart (2003) conducted a study with both applicants and incumbents using both methods to examine responses to SJTs, biodata, and personality measures. Ployhart et al. noted that within the applicant samples, the Web-based measures tended to be superior to paper-and pencil measures because the data tended to cluster around the mean, had lower means, higher internal consistency, more variance, and tended to be more highly correlated. Differences between Web-based and paper and pencil personality measures tended to be the largest; differences between paper-and-pencil and Web administration of SJT measure were only slightly smaller than for personality measures. These results indicated that in applicant settings, Web measures have more favorable psychometric properties. There are two notable differences between Ployhart's et al. research and the present study. First, the differences observed on SJT scores in Ployhart et al. may have resulted from the nature of the sample. Ployhart et al. suggested that applicants may have answered with what they thought was the best answer (i.e., should do), and incumbents may have answered with

what they have done in the past (i.e., would do). McDaniel et al.'s (2007) research supported that SJTs that utilize knowledge instructions (e.g., "should do") have higher positive correlations (.35) with cognitive ability than do SJTs that utilize behavioral instructions (.19) (e.g., "would do"). Behavioral instructions better correlate with personality constructs. McDaniel et al. concluded that when instructions ask a participant what they should do, participants look for the answer that is the best, or maximally correct, whereas participants who are asked what they would do in a particular situation look for the answer that best suits their personality.

In the present study, all participants will be from the same organizational level and will be asked what they should do, using the SJT as a measure of cognitive ability. Ployhart et al. used a personality-based SJT, whereas a cognitive ability SJT will be employed in the current research. A description of the SJT (i.e., SALSA©) used in the current study follows.

*SALSA©*

The SJT that is the focus of the current study is the Situational Assessment of Leadership – Student Assessment (SALSA©), which was developed to assess the seven most common leadership assessment center dimensions reported by Arthur, Day, McNelly, and Edens (2003): Organizing/Visioning/Planning; Consideration/Team Skills; Problem Solving/Innovation; Influencing Others; Communication; Drive/Results Orientation; and Tolerance for Stress. Additionally, another dimension, Integrity/Ethics was included. There are a total of 120 items across all eight dimensions. Students in an Industrial/Organizational (I/O) Psychology Masters program, Honors Leadership students, and members of The Dynamic Leadership Institute served as SMEs to generate

critical incidents as well as to provide three or four response alternatives for each situation (Grant, 2009). A scoring key was developed using a process consistent with Motowidlo et al. (1990) and Lievens et al. (2008). Seven university faculty members with considerable experience teaching leadership at the undergraduate and graduate level served as SMEs and rated the effectiveness of each response option. Only items that had one correct alternative, as indicated by SME ratings, were retained in the final version of SALSA©.

SALSA© asks participants to select the response for each item that represents the behavior they believe a leader should engage in for the most effective leadership response in the situation described; this is a measure of cognitive ability (McDaniel et al., 2007). Research indicates internal consistency of the SALSA© is $\alpha = .91$ (Grant, 2009). An analysis of the difficulty of SALSA© items indicated nearly an even number of easy, moderate, and difficult items (Grant). In addition, in pilot testing of SALSA© there were significant main effects found for gender ($M_{\text{Females}} = 82.30$, $SD = 14.44$; $M_{\text{Males}} = 72.41$; $SD = 15.04$), students whose primary language was not English ($M_{\text{English}} = 88.27$, $SD = 15.55$; $M_{\text{ESL}} = 65.67$; $SD = 19.75$), and gender for ESL students ($M_{\text{Females}} = 44.00$, $SD = 9.40$; $M_{\text{Males}} = 69.87$; $SD = 20.21$; Grant). Grant also examined the convergent validity between the Center for Leadership Excellence Assessment Center scores and SALSA© scores. Convergent validities for the matched dimensions ranged from $r = .28$ to $r = .44$, indicating low but significant correlations (Grant). As the current study will examine the convergent validity between SALSA© and rating from supervisors, peers, and self, the literature on multi-source feedback will be reviewed next.

*Multi-Source Feedback*

Supervisor ratings are a very common form of performance evaluation (Foster & Law, 2006; Van Hooft, Van der Flier, & Minne, 2006; Wilkerson, Manatt, Rogers, & Maughan, 2000). Ratee feedback typically is based on the behavior of the individual, and is often used in conjunction with an appraisal instrument that measures particular dimensions of behavior. Three-hundred-sixty degree and multi-source feedback expand the information available to an assessee and can be used for administrative, developmental, or research purposes (Van Hooft, et al.; Jackson & Greller, 1998). Three-hundred-sixty degree feedback and multi-source feedback are not synonymous terms. According to Foster and Law (2006), 360˚ feedback is a form of multi-source feedback. As the name indicates, 360˚ feedback consists of a full circle (thus the name 360˚ feedback) of raters from different organizational levels. These raters include superiors, peers, subordinates, and self. Multi-source ratings are defined as ratings from two or more raters who provide personalized feedback. A further distinction among raters is in terms of organizational level. Varying expert power (e.g., a tenured employee versus a newly hired employee that fulfills the same position) between raters does not represent different organization levels (Foster & Law). In other words, two employees varying in expert power in the same position cannot serve as two different levels of raters to satisfy the multi-source requirement; they must actually be from two different levels (i.e., superior and subordinate) in the organization.

Using multi-source feedback in practice requires important process implementation guidelines to achieve quality outcomes. For example, Antonioni (1994) suggested that raters should remain anonymous. In a multi-source survey conducted by

Antonioni, some subordinates remained anonymous, while others were identified. He acknowledged that those in the anonymous condition gave better quality ratings; that is, the scores were less inflated. Similarly, London and Wohlers (1991) asked participants in their study if they had to rate their boss and be held accountable whether their ratings would have been different. Nearly a quarter of the participants indicated they would have rated their boss differently. Although the managers in Antonioni's study stated they preferred ratings from identified sources and that they were more accepting of the results, anonymous ratings are recommended because ratings are more accurate.

Another implication for the quality of ratings is whether the feedback is used for developmental versus appraisal purposes. For example, in one study where students were on teams, those students who were told to rate others for appraisal purposes tended to assign nearly the same score on every item (halo effect) and were much more lenient than those who were told the ratings were strictly for developmental purposes (Farh, Cannella, & Bedeian, 1991). Similar to the effects of anonymous versus accountable ratings, 34% of raters indicated that if the feedback were to be used for performance appraisal versus developmental feedback their ratings would have differed. Dalessio (1998) suggested that rater errors can be reduced by providing rater training that can include information specifically related to rater errors, as well as information on the purpose of the process and how the instrument was designed. Additional suggestions included using feedback as a means of developmental improvement rather than appraisal, and prompting recipients of feedback to be more receptive to negative feedback.

Research also speaks to the effect of using multiple sources rather than a single source for feedback. Bernardin, Dahmus, and Redmon (1993) found that attitudes were

more positive about feedback when both managers and subordinates provided feedback than when just managers or just subordinates provided feedback. Interestingly, the results of this study imply that ratees are aware of the different results from ratings by different raters, and perhaps value the results more when they are more comprehensive.

Aside from what ratees merely think about the ratings they receive based on the sources and processes used to obtain ratings, research has indicated reliable differences between rater sources. For example, self-ratings tend to be higher than other-ratings (i.e., Bass & Yammarino, 1991; Brutus, Fleenor, & McCauley, 1999; Harris & Schaubroeck, 1988). In a study that examined Navy officers, the discrepancy between self- and other-ratings was related to the actual leadership success of the ratee (Bass & Yammarino). Those officers that were actually less successful often rated themselves much higher than those officers that were successful; thus the discrepancies between self- and other-ratings were much higher for those who were less successful. Van Velsor, Taylor, and Leslie (1993) found similar results where self-overraters received the lowest ratings from subordinates. Contrary to Bass and Yammarino, however, Van Velsor et al. and Fletcher and Baldry (2000) indicated that those who underrate themselves may have received high overall ratings from others as result of being seen as by others as highly self-aware of their own shortcomings. Thus, the ratings from others were high because other-raters felt that the ratee was aware of his/her inadequacies. Delessio (1998) suggested that seeking negative feedback may be most beneficial in receiving effective ratings from all sources and obtaining self-other agreement.

Lievens et al. (2005) acknowledged that inter-rater correlations are sometimes low because of differences in meaning of dimensions when asked to rate an individual's

performance. For example, if dimensions are not clearly defined and understood similarly by raters, the result is different ratings on the same individual not because of actual differences in viewed performance, but because of differences in dimension meaning. Similarly, McDaniel, et al. (2007), Weekley and Ployhart (2005), and Van Hooft et al. (2006), all noted that while low inter-rater correlations can be problematic for administrative purposes, they can actually be beneficial for developmental purposes. That is, different raters provide different perspectives, and as such, different developmental opportunities.

Little research on the relationship between multi-source ratings with external measures has been conducted. However, there have been attempts to examine construct validity of multi-source ratings by comparing the ratings within and between different sources (e.g., self-other agreement). Previous research has indicated a moderate positive relationship between averaged task ratings and self-, supervisors-, and peer ratings (Vance, Coovert, MacCallum, & Hedge, 1989; Lance, Teachout, & Donnelly, 1992). Nowack, Hartley, and Bradley (1999) and Lievens et al. (2005) both noted that different rater groups tend to focus on different dimensions of ratee performance. For example, bosses are more likely to focus on bottom-line performance, whereas co-workers (peers) put more emphasis on interpersonal and relationship factors. Other research has shown that supervisor-ratings are more reliable than ratings from other sources (Van Hooft et al., 2006; Weekley, Ployhart, & Harold, 2004). It may be that different types of raters do not have the same opportunity to observe behaviors reflecting all dimensions (Conway & Huffcutt, 1997). Accordingly, it becomes even more important to focus on specific dimensions that can be analyzed from multiple perspectives.

Research also has indicated that self-ratings tend to be higher than other-ratings (i.e., Bass & Yammarino, 1991; Brutus et al., 1997; Harris & Schaubroeck, 1988). Higher self-ratings are a result of making ourselves feel better about our behavior and in hopes that whoever is using the instrument will be tempted to agree with our self-assessment. Because SJTs are susceptible to distortions associated with self-deception and impression management, participants in the current study will be asked to rate themselves honestly, and self-ratings will be correlated with the ratings of others to check rater agreement (McDaniel et al., 2007; Weekley et al., 2004). Assessment center ratings have been used to validate 360˚ ratings, but most studies found non-significant results. Lievens et al. (2008) noted that SJTs are measurement methods for assessing a variety of constructs, and any correlation "with personality and cognitive ability depends on the constructs measured and on the response instructions used" (p. 431). In the current study, the SJT dimensions are a one-to-one match to the rated dimensions.

In terms of criterion-validity, Dalessio (1998) suggested that the source of the feedback should be related to the target criterion. For example, subordinate ratings may be used better to assess job satisfaction and turnover within the department, while supervisor ratings may best be used to evaluate production. Each respective source has information that is relevant to the respective criterion. Further, criterion-validity can be established for multi-source instruments if there is evidence that they distinguish between those performing effectively and poorly on the same criterion variable.

Research on the reliability of multi-source feedback indicates that internal consistency is reasonably good with coefficient alphas in the .70 range (Van Velsor & Leslie, 1991). Additionally, there seems to be moderate agreement within raters (e.g., all

peers) versus between raters (e.g., superiors and peers), but that agreement among supervisors (.51) and among peers (.39) is higher than that among subordinates (.27; Conway & Huffcutt, 1997). Agreement between rating sources may not be the most appropriate reliability estimate as different rating sources may differ in their access to ratee performance.

Most practitioners working with organizations want to know the benefits of implementing multi-source feedback and, more specifically, how it might improve performance. The good news is that it appears that even without training or development intervention, multi-source feedback improves performance. Hazucha, Hezlett, and Schneider (1993) conducted a study in which 48 managers were assessed on job-related dimensions on two different occasions, two years apart. Results indicated that other-ratings improved and self-other ratings were closer in agreement after two years. Atwater, Roush, and Fischthal (1995) found similar results when leaders at the U.S. Naval Academy were assessed on their performance; the ratings were more comparable during the second test administration, indicating that their behaviors had improved. In one study, however, increased ratings only appeared in subordinate, peer, and customer ratings; supervisor and self-ratings did not improve (Bernardin, Hagan, Ross, & Kane, 1995, as cited in Dalessio, 1998). Thus, even without intervention, it appears that self-awareness is a mediator for behavior improvement. When employees are provided with performance feedback, they become more aware of desired behaviors and strive to achieve them. This change in behavior is seen by subordinates and peers.

The previous research begs the question of whether feedback is necessary for performance improvement. Dominick, Reilly, and McGourty (1997) concluded that

feedback is not necessary. Their study examined a group of teams that were divided into groups that either simply rated themselves and others, but received no feedback (exposure group), rated themselves and others and received feedback (feedback group), or not perform any ratings (control group). While both the feedback group and exposure group had higher assessor ratings than the control, there were no differences between the exposure and feedback group. Again, self-awareness played an important role, but multi-source instruments also seemed to play a valuable role in indicating what knowledge, skills, and abilities an organization values, providing employees with helpful information to guide their behavior without formalized feedback. Performance improvement through mere implementation of multi-source feedback may have utility for organizations because at the very least, multi-source feedback can yield positive results without the added costliness of training and development that feedback may require. Still, there is substantial research that indicates feedback, in relation to goals, is a necessary mediator to change performance. Locke and Latham (2002) argued that in order for goals to be effective, feedback must be given in relation to those goals. Feedback provides important information that serves both a cueing and a motivational function. In the context of performance ratings, Locke and Latham would suggest that without feedback from others, people merely have an idea of what is expected of them (that is, if they completed a self-ratings form) or a goal, but no information on where they stand in relation to that goal.

Research supports that SJTs and multi-source rating feedback are both relatively popular methods of evaluation, and for good reason. Both provide valuable information about the participant. SJTs have the potential to yield personality or cognitive

information; the multi-source ratings serve as a gauge of current performance. Additionally, both measures are versatile in that a variety of constructs can be measured. Finally, both are psychometrically sound methods for assessing performance.

*The Current Research*

The current research assessed the convergent validity of multi-source ratings with SALSA© scores. Army Military Science cadets completed SALSA©, and rated themselves on the eight leadership dimensions. Ratings also were provided by one superior and multiple peers. SALSA© scores and rating feedback were provided to the participants and superiors for potential developmental purposes; feedback will not be used for administrative purposes.

Despite that Dominick et al. (1997) concluded that feedback is not necessary for performance improvement, Jackson and Greller (1998) cautioned that data, evaluation, and action are all necessary components of feedback. Furthermore, there is substantial research indicating feedback is a necessary element for goals to influence performance (Locke & Latham, 2002). As such, cadets received feedback on the eight dimensions of leadership ratings identified by each source and on SALSA©. The feedback indicated where the cadet fell relative to other cadets on each measure. The supervising officer also received the results for his/her cadets and information concerning the relationship between SALSA© scores, peer ratings, cadre ratings, and self-ratings.

In the current study, analyses included correlations between overall SALSA© scores with overall ratings from each source; individual SALSA© dimension scores were correlated with ratings for each dimension.

*Hypotheses*

The following hypotheses were tested. Previous research (McDaniel et al., 2007; Weekley & Ployhart, 2005) has indicated that cognitive ability is at least moderately correlated with SJT scores. As SALSA© asks participants what *should* be done in each scenario (i.e., SALSA© is testing cognitive ability), the first hypothesis is as follows.

Hypothesis 1: Overall SALSA© scores will correlate with overall performance ratings on Problem Solving/Innovation.

The eight dimensions on the rater feedback forms parallel the eight dimensions of SALSA©; it was expected that each score for each dimension in SALSA© would positively correlate with the corresponding dimension on the rating form.

Hypotheses 2a-h: Each dimension score on SALSA© will positively correlate with the corresponding performance rating.

Hypothesis 2a: SALSA© dimension Organizing/Visioning/Planning will positively correlate with performance rating dimension Organizing/Visioning/Planning.

Hypothesis 2b: SALSA© dimension Consideration/Team Skills will positively correlate with performance rating dimension Consideration/Team Skills.

Hypothesis 2c: SALSA© dimension Problem Solving/Innovation will positively correlate with performance rating dimension Problem Solving/Innovation.

Hypothesis 2d: SALSA© dimension Influencing Others will positively correlate with performance rating dimension Influencing Others.

Hypothesis 2e: SALSA© dimension Communication will positively correlate with performance rating dimension Communication.

Hypothesis 2f: SALSA© dimension Drive/Result Orientation will positively correlate with performance rating dimension Drive/Result Orientation.

Hypothesis2g: SALSA© dimension Tolerance for Stress will positively correlate with performance rating dimension Tolerance for Stress.

Hypothesis2h: SALSA© dimension Integrity/Ethics will positively correlate with performance rating dimension Integrity/Ethics.

Hypotheses 3a-h are based on previous research that has indicated that self-ratings tend to be higher than ratings from other sources (Bass & Yammarino, 1991; Brutus et al., 1997; Harris & Schaubroeck, 1988).

Hypotheses 3a-h: Self-ratings will be higher than ratings from other sources on all eight SALSA© dimensions.

Hypothesis 3a: Self-ratings will be higher than other-ratings on SALSA© dimension Organizing/Visioning/Planning and performance rating dimension Organizing/Visioning/Planning.

Hypothesis 3b: Self-ratings will be higher than other-ratings on SALSA© dimension Consideration/Team Skills and performance rating dimension Consideration/Team Skills.

Hypothesis 3c: Self-ratings will be higher than other-ratings on SALSA© dimension Problem Solving/Innovation and performance rating dimension Problem Solving/Innovation.

Hypothesis 3d: Self-ratings will be higher than other-ratings on SALSA© dimension Influencing Others and performance rating dimension Influencing Others.

Hypothesis 3e: Self-ratings will be higher than other-ratings on SALSA© dimension Communication and performance rating dimension Communication.

Hypothesis 3f: Self-ratings will be higher than other-ratings on SALSA© dimension Drive/Result Orientation and performance rating dimension Drive/Result Orientation.

Hypothesis 3g: Self-ratings will be higher than other-ratings on SALSA© dimension Tolerance for Stress and performance rating dimension Tolerance for Stress.

Hypothesis 3h: Self-ratings will be higher than other-ratings on SALSA© dimension Integrity/Ethics will and performance rating dimension Integrity/Ethics.

Other research (Weekley, et al., 2004) has demonstrated that employment status affects SJT performance. For example, incumbents scored higher than applicants. This is

assumed to be a function of cognitive ability resulting from more time on the job. Hypothesis 4 follows:

Hypothesis 4a: Seniors will perform better on the SJT than will Juniors.

Hypothesis 4b: Seniors will receive higher performance ratings than will Juniors.

## Method

### *Participants*

Forty-one Western Kentucky University ROTC students participated in the study. However, 7 participants were removed from the sample for one of three reasons: 2 participants did not attempt SALSA©; 4 participants started, but did not complete SALSA©; and one participant's SALSA© score was below what would be expected by chance and, as such, strongly suggested the student was inattentive when completing SALSA©. Thus, data from a total of 34 participants were analyzed. Participants were either Seniors (N = 8) or Juniors (N = 26) in the ROTC. There were 31 males and 3 females, with an average age of 23.00 years (*SD* = 4.03).

### *Materials and Procedure*

SALSA© was administered online via Blackboard, a Web-based course-management system designed to allow students and faculty to participate in classes delivered online or to use online materials and activities. Cadets were instructed to go on-line to complete SALSA©. Completion of SALSA© took approximately one hour. SALSA© consists of 130 items across eight subtests, each measuring a different dimension of leadership. Each test item presents a realistic but hypothetical leadership scenario and asks participants to select from four multiple-choice response options the behavior they believe a leader should engage in for the most effective leadership response.

The cadets participated in a multi-source feedback process in which they were rated on eight dimensions of leadership by three types of raters: his/herself, cadre officer(s), and cadet peers. The rating forms were consistent with SALSA© in that raters

were provided with a description of each of the eight leadership dimensions and were asked to rate each cadet (or his- or herself) on a 5-point scale. The cadets were provided with a packet that contained one self-rating form and rating forms identified for 10 randomly assigned cadets from their class (i.e., Juniors rated Juniors and Seniors rated Seniors). Each participant was asked to complete the self-rating form and to rate the 10 identified cadets within two weeks of receiving the packet. Thus, each cadet could have been rated by up to 10 peers. ROTC Cadre received a rating form for each of the cadets in their class. Juniors were rated by one cadre member and seniors were rated by two cadre members.

## Results

An overall SALSA© score was calculated for each individual by summing across the eight dimension scores. Two overall peer ratings were calculated for each cadet by averaging ratings from each peer across dimensions and by averaging the peer Overall Leadership Excellence ratings. Two overall cadre ratings were calculated for each cadet by averaging ratings from each cadre across dimensions as well as averaging the cadre Overall Leadership Excellence ratings. For peer/cadet ratings, the averaged ratings across dimensions ($M$ = 3.82, $SD$ = .47) and the Overall Leadership Excellence ratings ($M$ = 3.82, $SD$ = .53) were equal. For cadre, the averaged ratings across dimensions ($M$ = 3.68, $SD$ = .92) and the Overall Leadership Excellence ratings ($M$ = 3.60, $SD$ = .74) were not significantly different ($t(33) = -.60$, $p = .55$). As such, the Overall Leadership Excellence ratings were used for further analyses. T-tests to determine whether the cadre ratings for Juniors and Seniors differed indicated that Junior cadre ratings ($M$ = 3.57, $SD$ = .70) did not differ significantly from Senior cadre ratings ($M$ = 3.68, $SD$ = .92), $t(25) = -.47$, $p = .46$).

Overall Leadership Excellence (OLE) ratings from each source (i.e., peer, cadre, and self) were correlated with the Total SALSA© score. None of the resulting correlations were significant. Specifically, the correlation between Total SALSA© score and OLE peer ratings was not significant ($r = .14$, $p = .42$), nor was the correlation between Total SALSA© score and OLE cadre ratings ($r = .20$, $p = .24$). Similarly, the correlation between Total SALSA© score and OLE self-ratings was not significant ($r = -.12$, $p = .47$).

Hypothesis 1, which predicted that Total SALSA© score would positively correlate with the Problem Solving/Innovation ratings, was not supported for any of the ratings (self ($r$ = .10, $p$ = .54), cadre ($r$ = .08, $p$ = .65), and peer ratings ($r$ = .08, $p$ = .64)).

Hypotheses 2a-h, which predicted that each dimension score on SALSA© would positively correlate with the corresponding performance rating, was tested by correlating each SALSA© dimensions score with the corresponding dimension rating from each rating source. Results indicated a significant correlation between the Consideration/Team Skills SALSA© score and the cadre rating for the same dimension ($r$ = .44, $p$ < .01), supporting Hypothesis 2b. No other SALSA© dimension scores correlated with any of the multi-source ratings of the same dimension. Thus, only one of Hypotheses 2a-h received support (i.e., only for Hypothesis 2b). It is of interest to note that for all of the dimensions, cadre and peer ratings were significantly correlated; likewise, cadre and peer Overall Leadership Excellence (OLE) ratings were significantly correlated ($r$ = .72, $p$ < .01). There were two dimensions (i.e., Problem Solving/Innovation and Drive/Results Orientation) for which there were significant correlations between cadre ratings and self-ratings. See Table 1 for the correlation matrix between specific rating sources.

Hypothesis 3a-h stated that self-ratings on each of the eight leadership dimensions would be higher than ratings by cadets/peers and cadre. These hypotheses were tested by independent sample t-tests. Overall self-ratings were significantly higher ($M$ = 4.08, $SD$ = .45) than overall cadre ratings ($M$ = 3.60, $SD$ = .74; $t$(33) = -2.76, $p$ < .01), and significantly higher than overall peer ratings ($M$ = 3.82, $SD$ = .53; $t$(33) = -3.72, $p$ < .01). Self-rating for each of the eight leadership dimensions were significantly higher than both cadre and peer ratings for each of the dimensions, respectively (see Table 2).

Table 1

*Correlations Between Ratings from Different Sources on SALSA© Dimensions*

| | Self | Peer | Cadre |
|---|---|---|---|
| Organizing/Planning/Visioning | | | |
| Self | -- | 0.22 | 0.13 |
| Peer | | -- | .64** |
| Cadre | | | -- |
| Consideration/Team Skills | | | |
| Self | -- | 0.18 | 0.24 |
| Peer | | -- | .51** |
| Cadre | | | -- |
| Problem Solving/Innovation | | | |
| Self | -- | -.43* | -.36* |
| Peer | | -- | .68** |
| Cadre | | | -- |
| Influencing Others | | | |
| Self | -- | 0.02 | 0.14 |
| Peer | | -- | .61** |
| Cadre | | | -- |
| Communication | | | |
| Self | -- | 0.21 | 0.14 |
| Peer | | -- | .42* |
| Cadre | | | -- |
| Drive/Results Orientation | | | |
| Self | -- | 0.23 | .36* |
| Peer | | -- | .73** |
| Cadre | | | -- |
| Tolerance for Stress | | | |
| Self | -- | 0.09 | 0.17 |
| Peer | | -- | .62** |
| Cadre | | | -- |
| Integrity/Ethics | | | |
| Self | -- | -0.06 | -0.13 |
| Peer | | -- | .41* |
| Cadre | | | -- |

*$p < .05$, **$p < .01$

Table 2

*Results for Self-Ratings Compared to Peer and Cadre Ratings for 8 Dimensions of Leadership*

| Dimension | *M* | *SD* | *df* | *t* |
|---|---|---|---|---|
| Organizing/Planning/Visioning | | | | |
| Self | 4.09 | 0.52 | - | - |
| Peer | 3.87 | 0.57 | 33 | -2.14* |
| Cadre | 3.69 | 0.85 | 33 | -2.72* |
| Consideration/Team Skills | | | | |
| Self | 4.32 | 0.63 | - | - |
| Peer | 3.84 | 0.53 | 33 | -5.17** |
| Cadre | 3.52 | 0.72 | 33 | -6.33** |
| Problem Solving/Innovation | | | | |
| Self | 4.08 | 0.66 | - | - |
| Peer | 3.77 | 0.52 | 33 | -3.40** |
| Cadre | 3.83 | 0.75 | 33 | -1.86* |
| Influencing Others | | | | |
| Self | 3.97 | 0.67 | - | - |
| Peer | 3.60 | 0.57 | 33 | -3.75** |
| Cadre | 3.45 | 0.78 | 33 | -3.83** |
| Communication | | | | |
| Self | 4.23 | 0.78 | - | - |
| Peer | 3.84 | 0.54 | 33 | -4.13** |
| Cadre | 3.72 | 0.75 | 33 | -3.95** |
| Drive/Results Orientation | | | | |
| Self | 4.14 | 0.70 | - | - |
| Peer | 3.82 | 0.63 | 33 | -2.87** |
| Cadre | 3.58 | 0.90 | 33 | -3.57** |
| Tolerance for Stress | | | | |
| Self | 3.91 | 0.79 | - | - |
| Peer | 3.71 | 0.50 | 33 | -2.28* |
| Cadre | 3.57 | 0.79 | 33 | -2.45* |
| Integrity/Ethics | | | | |
| Self | 4.52 | 0.56 | - | - |
| Peer | 4.15 | 0.48 | 33 | -4.35** |
| Cadre | 4.08 | 0.62 | 33 | -4.05** |
| Overall Leadership Excellence | | | | |
| Self | 4.08 | 0.45 | - | - |
| Peer | 3.82 | 0.53 | 33 | -3.72** |
| Cadre | 3.60 | 0.74 | 33 | -2.76** |

Note: t-test compares to self-rating.
*$p < .05$, **$p < .01$

Hypothesis 4a stated that Seniors would perform better on the SJT than would Juniors and Hypothesis 4b stated that Seniors would receive higher performance ratings than would Juniors. An independent samples t-test indicated that Junior SALSA© scores ($M$ = 84.42, $SD$ = 8.24) were significantly lower than those of Seniors (M = 88.87, $SD$ = 10.11), $t(25) = -2.74$, $p = .01$. Thus, Hypothesis 4a was supported. However, Seniors ($M$ = 3.68, $SD$ = .92) were not rated significantly higher than Juniors ($M$ = 3.57, $SD$ = .70) by their respective cadre, $t(25) = -.748$, $p = .46$. Likewise, Seniors ($M$ = 3.90, $SD$ = .85) were not rated significantly higher than Juniors ($M$ = 3.80, $SD$ = .40) by their respective peers, $t(25) = -1.21$, $p = .23$. Thus, Hypothesis 4b was not supported.

## Discussion

### *Convergent Validity*

The purpose of this study was to compare multi-source ratings and SALSA© scores to examine convergent validity. Overall ratings were correlated with overall SALSA© scores, and individual ratings on dimensions were correlated with corresponding dimensions on SALSA©. There were no significant correlations found in any of the analyses, with one exception. Cadre ratings of Consideration/Team Skills were positively correlated with SALSA© scores on the same dimension. These findings suggest that the multi-source ratings and the SJT are not measuring the same leadership construct. With one exception, the SJT scores are independent of the ratings provided by the ROTC members.

The participants used in this research were military leaders, which may provide some insight into the non-significant correlations. SALSA© was keyed based on a participative, democratic model of leadership. The military model of leadership is much more autocratic. Interestingly, the only dimension on which cadre ratings correlated with SALSA© scores was Consideration/Team Skills, the dimension that emphasizes rapport, respect, and two-way communication. Military personnel are often taught to follow protocol for many situations that may not fit the given response options offered on SALSA©. Former Secretary of Defense William Cohen once stated, “One of the challenges for me is to somehow prevent a chasm from developing between the military and civilian worlds where the civilian world doesn’t fully grasp the mission of the military, and the military doesn’t understand why the memories of our citizens and civilian policy-makers are so short, or why the criticism is so quick and so unrelenting”

(as cited in Allen & Coates, 2010, p. 75). For example, in a scenario where a civilian manager witnesses another manager stealing, it may be appropriate for the manager to confront his fellow peer; but in the military, two leaders with equal ranking in the same situation are not be permitted to confront each other and, instead, are required to report to the chain of hierarchy. This is a simplistic example of how military and civilians may judge the correct response differently in similar situations. Another difference between military leaders and civilian leaders is what is being taught as important characteristics of leadership. For example, there is a current push for military leaders to be cross-culturally educated and aware (Abbe & Halpin, 2010), a dimension that is not measured by SALSA©.

*Additional Findings*

Consistent with previous research (i.e., Bass & Yammarino, 1991; Brutus, Fleenor, & McCauley, 1999; Harris & Schaubroeck, 1988), when self-ratings were compared to the ratings from other sources, self-ratings were higher than those from either cadre or peers on all eight dimensions of leadership as well as on Overall Leadership Excellence. This finding indicates that cadets who rated themselves may have used impression management tactics (i.e., inflated ratings) while responding to the self-assessment, despite being asked to rate themselves honestly.

The fact that Seniors significantly outperformed Juniors on SALSA© comes as no surprise as previous research (Grant, 2009) revealed a potential experience effect in which graduate students with more education outperformed undergraduate students. These findings suggest that some graduate coursework or, in this case, more

undergraduate education, particularly in leadership, may provide a better understanding of the organizational situations represented in SALSA©.

Interestingly, Seniors did not receive higher performance ratings than Juniors. One would expect the group dynamic among Seniors to be stronger than the bond between Juniors and, further, that Seniors would rate their fellow peers higher because they have had time to build a better interpersonal bond. However, the current study did not yield such results. With regard to the cadre ratings, the fact that cadre rated only one class of cadets (i.e., either Juniors or Seniors) may have contributed to a lack of difference between the two groups. That is, a single cadre rated his cadets without information about their performance relative to that of the other class of cadets.

Another finding of this research that was not predicted and even contradicted previous research (Conway & Huffcutt, 1997) was that peers and superiors significantly and positively agreed in their rating of the target individual. McDaniel et al. (2007), Weekley and Ployhart (2005), and Van Hooft et al. (2006) indicated that different rating sources may differ in their access to ratee performance and subsequently provide data that differ. It was expected that the cadre would focus on bottom-line performance, whereas other cadets (peers) would put more emphasis on interpersonal and relationship factors when they considered the target individual they were rating (Nowack et al., 1999; Lievens et al., 2005). However, the current research found that superiors and peers rated the cadets similarly.

Self-ratings and cadre ratings were correlated on two dimensions; the ratings on the Drive/Results Orientation dimension were positively correlated and ratings on the Problem Solving/Innovation were negatively correlated. It is unclear why self-ratings and

cadre ratings were positively correlated on only one dimension. It may be possible that in most civilian leadership situations, superiors and peers evaluate leadership differently (e.g., McDaniel et al., 2007; Weekley & Ployhart, 2005). In military situations, perhaps the degree to which an individual originates and maintains a high activity level, sets high performance standards and persists in achievement, and expresses the desire to advance to higher job levels is more evident, resulting in similar ratings. Military personnel are taught early in training the importance of clear direction, to push one's self and others for high quality and results, to monitor progress and results, and to demonstrate a bias for action. Thus, the Drive/Results Orientation dimension may be viewed similarly by those rating themselves and by their superiors. On the other hand, the negative correlation on the Problem Solving/Innovation dimension may be a result of self-raters giving themselves higher ratings than deserved, whereas cadre may accurately recognize their weak performance and rate the cadets accordingly, or conversely, the cadet may be providing accurate self-ratings on this dimension while cadre ratings are overly harsh.

*Limitations*

There were several limitations to the current study. A potential limitation involves the development of SALSA©. If, in fact, military and civilian leaders differ in effective leadership responses, it may be that SALSA© specifically targets what civilian leaders should do rather than what military leaders should do. Given that SALSA© was keyed toward a participative, democratic leadership style, it may be that for it to be used effectively in autocratic situations (e.g., military), the scenarios and options need to be revised to better match autocratic leadership model.

Given that this validation study may have revealed potential differences in military versus civilian leadership, a similar study should be conducted to test the reliability of these findings. It is recommended that a similar study be conducted with both a new sample of cadets and a sample of civilian leaders (e.g., Leadership Certificate students). Additional data on SALSA© may reveal leadership differences among civilians versus military personnel, programs, and experience or education. Further, the validity of the SJT may increase as sample size increases with more raters, as a sample size of 34 is small.

Another limitation to the study is the time frame in which the SALSA© and multi-source ratings were completed. Raters were asked to complete both the SALSA© and the rating forms within a two week time period, and to complete all of the rating forms in one sitting. It is unclear whether these directions were followed. A break in completing the rating forms may have resulted in rater errors. For example, if a participant rated some of the target cadets at one time period, took a break, and resumed rating others later, his/her frame of reference may have changed for a variety of reasons (e.g., a change in demeanor or having a recent interaction with someone), resulting in rating error.

Although frame of reference is one concern, there are other rating errors that only rater training can mitigate. For example, there may have been contrast effects (the rater evaluates the target individual relative to others rather than relative to standards), halo effects (generalizing one aspect of a person's behavior to all aspects of his/her behavior), similar-to-me effects (a rater rating a similar target more favorably than one who is

dissimilar to the rater), central tendency (consistently rating others at the midpoint of the scale), and positive or negative leniency (inaccurately rating others either high or low).

*Directions for Future Research*

Grant (2009) found that 46% of variance for those whose primary language was not English was accounted for by language differences. Specifically, students who spoke English as their primary language outperformed students who spoke English as a second language. Given that SALSA© may not measure leadership ability equally for all students, a demographic question about language should be added.

Future research could examine the possibility of a training effect. Thus, students from a variety of levels in leadership courses could be examined. Thus far, SALSA© seems to distinguish between those that have more leadership education (i.e., score higher) compared to those who have less education (i.e., score lower). Researchers should continue to examine this dynamic to determine the reliability of this finding. Again, the type of program in which a participant is involved may have significant implications for SALSA© scores and the ratings given by others (as witnessed with the ROTC program).

Finally, it might be helpful to examine the grades of participants. Both grades in leadership classes and overall GPA might serve as criterion variables to validate the SALSA©. This comparison would help to determine if SALSA© is indeed measuring leadership or some other construct such as general mental ability.

*Conclusions*

In sum, convergent validity between the SALSA©, an SJT, and multi-source ratings was examine. ROTC Cadets were asked to rate themselves and their peers on eight leadership dimensions. Cadre were asked to rate their respective subordinates.

Mostly nonsignificant correlations suggest that perhaps there are differences between what constitutes effective leadership for military and civilian leaders. Other findings were consistent with previous findings. For example, previous research (i.e., Bass & Yammarino, 1991; Brutus et al., 1999; Harris & Schaubroeck, 1988) revealed that self-ratings tend to be higher than other ratings. Additionally, this research also supports the potential program effect (i.e., students with more education perform better than those with less education) found in previous SALSA© research (Grant, 2009).

# References

Abbe, A., & Halpin, S. M. (2010, Winter). The cultural imperative for professional military education and leader development. *Parameters,* 20-31.

Allen, C. D., & Coates, B. E. (2010, Winter). The engagement of military voice. *Parameters*, 73-87.

Antonioni, D. (1994). The effects of feedback accountability on upward appraisal ratings. *Personnel Psychology*, *47,* 349-356.

Arthur, Jr., W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *53,* 125-154.

Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self- and follower ratings of leadership. *Personnel Psychology*, *34,* 251-280.

Bass, B. M., & Yammarino, F. J. (1991). Congruence of self and others' leadership ratings of naval officers for understanding successful performance. *Applied Psychology*, *40,* 437-454.

Bernardin, H. J., Dahmus, S. A., & Redmon, G. (1993). Attitudes of first-line supervisors toward subordinate appraisals. *Human Resource Management*, *32,* 315-324.

Brutus, S., Fleenor, J. W., & McCauley, C. D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Management Development*, *18,* 417-435.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, *10,* 331-361.

Dalessio, A. T. (1998). Using multisource feedback for employee development and personnel decisions. In Smither, J.W. (Eds.), *Performance appraisal: State of the art in practice* (pp.279-330). Jossey-Bass Publishers, San Francisco, CA.

Dominick, P. G., Reilly, R. R., & McGourty, J. W. (1997). The effects of peer feedback on team member behavior. *Group and Organization Management*, *22,* 508-520.

Farh, J. L., Cannella, A. A., Jr., & Bedeian, A. G. (1991). Peer ratings: The impact of purpose on rating quality and user acceptance. *Group and Organization Studies*, *16,* 367-386.

Fletcher, C., & Baldry, C. (2000). A study of individual differences and self-awareness in the context of multi-source feedback. *Journal of Occupational and Organizational Psychology, 73,* 303-319.

Foster, C. A., & Law, M. R. F. (2006). How many perspectives provide a compass? Differentiating 360-degree and multi-source feedback. *International Journal of Selection and Assessment*, *14,* 288-291.

Grant, K. L. (2009). The validation of a situational judgment test to measure leadership behavior. (Master's Thesis). Retrieved from TopSCHOLAR. Paper 64.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, *41,* 43-62.

Hazucha, J. F., Hezlett, S. A., & Schneider, R. J. (1993). The impact of 360-degree feedback on management skills development. *Human Resource Management*, *32,* 325-351.

Jackson, J. H., & Greller, M. M. (1998). Decision elements for using 360˚ feedback. *Human Resource Planning*, *21,* 18-28.

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77,* 437-452.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, *90,* 442-452.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of literature. *Personnel Review*, *37,* 426-441.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*, 705-717.

London, M., & Wohlers, A. J. (1991). Agreement between subordinate and self-rating in upward feedback. *Personnel Psychology*, *44,* 375-390.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60,* 63-91.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75,* 640-647.

Nowack, K. M., Hartley, J., & Bradley, W. (1999). How to evaluate your 360 feedback efforts. *Training & Development*, 48-53.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection & Assessment*, *11,* 1-16.

Shoenfelt, E. L. (2008). Situational Assessment of Leadership – Student Assessment© (SALSA©). Bowling Green, KY.

Van Hooft, E. A. J., Van der Flier, H., & Minne, M. R. (2006). Construct validity of multi-source performance ratings: An examination of the relationship of self-, supervisor-, and peer-ratings with cognitive and personality measures. *International Journal of Selection and Assessment*, *14*, 67-81.

Vance, R. J., Coovert, M. D., MacCallum, R. C., & Hedge, J. W. (1989). Construct models of task performance. *Journal of Applied Psychology*, *74*, 447-454.

Van Velsor, E., & Leslie, J. B. (1991). *Feedback to managers: Vol. 1. A guide to evaluating multi-rater feedback instruments.* Greensboro, NC: Center for Creative Leadership.

Van Velsor, E., Taylor, S., & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management*, *32*, 249-263.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, *52*, 679-700.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance*, *18,* 81-104.

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of

validity, measurement, and subgroup differences. *Human Performance*, *17,* 433-461.

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360˚ Feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*, 179-192.

# Appendix A

## HSRB Approval Form

A LEADING AMERICAN UNIVERSITY WITH INTERNATIONAL REACH
HUMAN SUBJECTS REVIEW BOARD

In future correspondence, please refer to HS10-113, December 2, 2009

Heather Strope-Smith
c/o Dr. Shoenfelt
Psychology
WKU

Heather Strope-Smith
& Dr. Shoenfelt:

Your research project, *Multi-source feedback in relation to the Situational Assessment of Leadership: Student Assessment (SALSA)*,was reviewed by the HSRB and it has been determined that risks to subjects are: (1) minimized and reasonable; and that (2) research procedures are consistent with a sound research design and do not expose the subjects to unnecessary risk. Reviewers determined that: (1) benefits to subjects are considered along with the importance of the topic and that outcomes are reasonable; (2) selection of subjects is equitable; and (3) the purposes of the research and the research setting is amenable to subjects' welfare and producing desired outcomes; that indications of coercion or prejudice are absent, and that participation is clearly voluntary.

1. In addition, the IRB found that you need to orient participants as follows: (1) signed informed consent is not required; (2) Provision is made for collecting, using and storing data in a manner that protects the safety and privacy of the subjects and the confidentiality of the data. (3) Appropriate safeguards are included to protect the rights and welfare of the subjects.

   **This project is therefore approved at the Expedited Review Level until November 30, 2010.**

2. Please note that the institution is not responsible for any actions regarding this protocol before approval. If you expand the project at a later date to use other instruments please re-apply. Copies of your request for human subjects review, your application, and this approval, are maintained in the Office of Sponsored Programs at the above address. Please report any changes to this approved protocol to this office. A Continuing Review protocol will be sent to you in the future to determine the status of the project. Also, please use the stamped approval forms to assure participants of compliance with The Office of Human Research Protections regulations.

Sincerely,

Paul J. Mooney, M.S.T.M.
Compliance Coordinator
Office of Sponsored Programs
Western Kentucky University

HUMAN SUBJECTS REVIEW BOARD WKU APPROVED

HSRB APPLICATION # 10-113
APPROVED 12/2/09 to 11/30/10
EXEMPT EXPEDITED FULL BOARD
DATE APPROVED 12/2/09

cc: HS file number Strope-Smith HS10-113